# Integration of genomic data into electronic health records

Daniel Masys, MD
Affiliate Professor
Biomedical & Health Informatics
University of Washington, Seattle

Major portion of today's lecture is based on this article, currently in press and available online at http://www.ncbi.nlm.nih.gov/pubmed/22223081

ARTICLE IN PRESS

Journal of Biomedical Informatics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Technical desiderata for the integration of genomic data into Electronic Health Records

Daniel R. Masys [a,*], Gail P. Jarvik [b,c], Neil F. Abernethy [a], Nicholas R. Anderson [a], George J. Papanicolaou [d], Dina N. Paltoo [e], Mark A. Hoffman [f], Isaac S. Kohane [g], Howard P. Levy [h]

[a] Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195-7240, United States
[b] Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195-7720, United States
[c] Department of Genome Sciences, University of Washington, Seattle, WA 98195-7720, United States
[d] Division of Prevention and Population Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, United States
[e] Advanced Technologies and Surgery Branch, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, United States
[f] Cerner Corporation, Kansas City, MO, United States
[g] Harvard-MIT Division of Health Sciences and Technology, Bioinformatics & Integrative Genomics, Cambridge, MA, United States
[h] Division of General Internal Medicine and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, United States

# Topics

- Context: Systems design issues in Healthcare
- Functional characteristics of an ideal system
- A prototype operational EHR with genomic decision support

# Systems Design Issues in Healthcare

- Current practice largely depends upon the clinical decision making ***capacity*** and ***reliability*** of autonomous individual practitioners, for classes of problems that routinely exceed the bounds of unaided human cognition

# Company announces low-cost DNA decoding machine

Updated 5d 11h ago



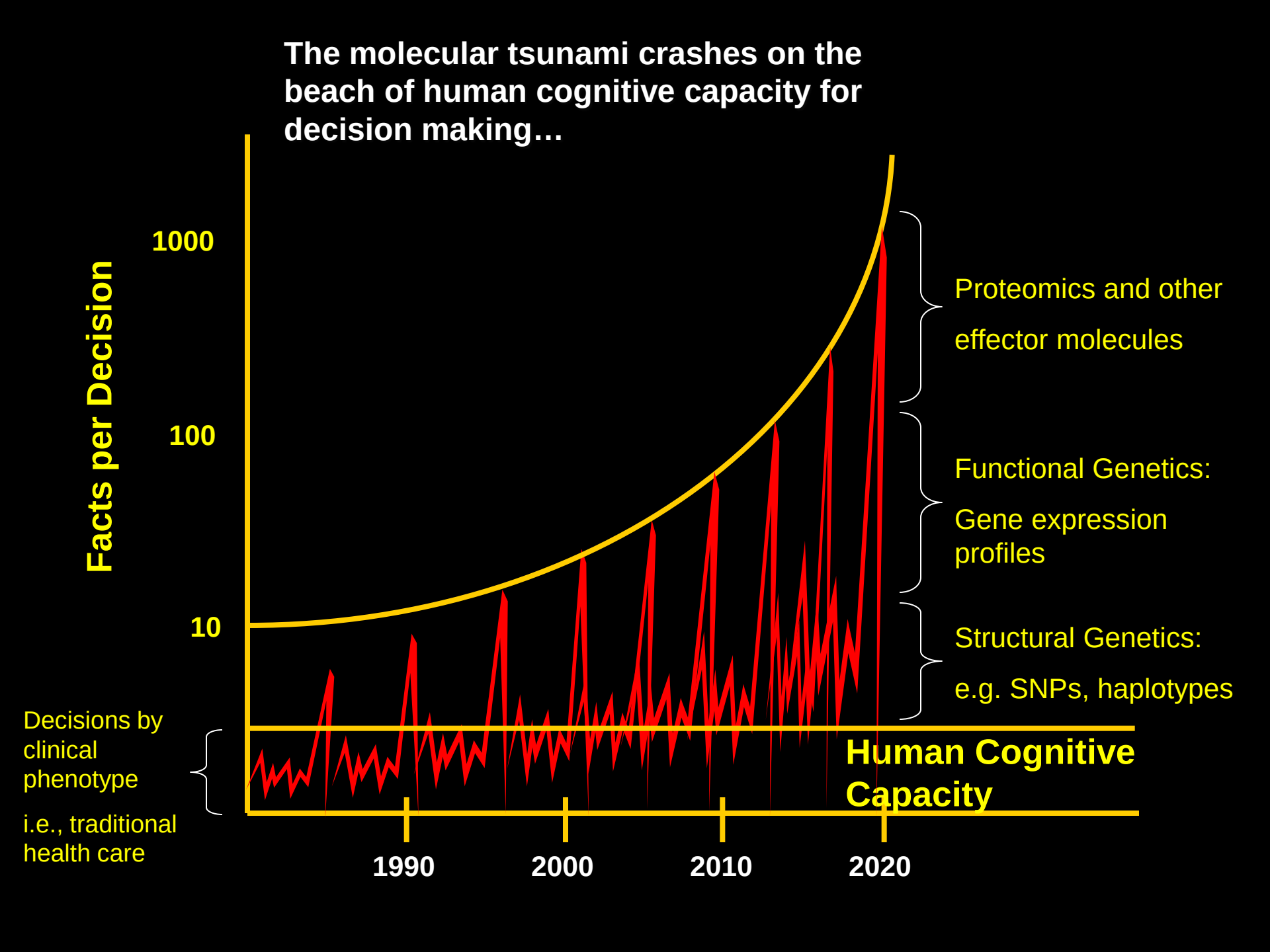
NEW YORK – A biotechnology company announced it has developed a machine to decode a person's DNA in a day for $1,000, a long-sought price goal for making a person's genome useful for medical care.

PR NEWSWIRE

[The Ion Proton(TM) Sequencer] Life Technologies Corp designed a machine to decode a person's DNA in a day for $1,000.

Life Technologies Corp. said Tuesday it was taking orders for the technology, which it expects to deliver in about a year. The Carlsbad, Calif., company said three major research institutions had already signed up for the $149,000 machine: the Baylor College of Medicine, the Yale School of Medicine and the Broad Institute of Cambridge, Mass.

The machine is a sequencer, meaning that it lets scientists identify the sequence of the 3 billion chemical building blocks that make up a person's DNA. Since the first sequencing of the basic human genome was announced at the White House in 2000, the costs of sequencing DNA have steadily tumbled. The $1,000 target has long been cited as a key step toward making the technique practical for doctors to use to help their


The molecular tsunami crashes on the beach of human cognitive capacity for decision making…
Facts per Decision
1000
100
10
Proteomics and other effector molecules
Functional Genetics: Gene expression profiles
Structural Genetics: e.g. SNPs, haplotypes
Decisions by clinical phenotype
i.e., traditional health care
Human Cognitive Capacity
1990
2000
2010
2020

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.

   **Issue:** current practice by clinical laboratories extinguishes many observations not felt to be clinically relevant.

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
   **Issue:** Methods used to perform genetic assays normally embedded in PDF text report but not available as structured data.

Firefox

Logical Observation Identifiers Names a...

loinc.org

Google

Home Downloads Documentation News About

# LOINC®

*Logical Observation Identifiers Names and Codes*

search

all of loinc.org

log in new user?

you are here: home

## Logical Observation Identifiers Names and Codes (LOINC®)

A universal code system for identifying laboratory and clinical observations.


Download Now (free)
Get started with LOINC today!

From serum levels of hepatitis B surface antigen to diastolic blood pressure, LOINC has standardized terms for all kinds of observations and measurements that **enable exchange and aggregation of electronic health data from many independent systems.**

> *More than 14,000 people in 145 countries use LOINC to help make bridges across their islands of health data.*

It's free, but invaluable. Both LOINC and the RELMA mapping program that helps link your local codes to LOINC terms are **distributed at no cost** by the Regenstrief Institute. LOINC is your key to interoperable data exchange.

### Ready to get started?

Or, search LOINC with our online app

**Learn**
**LOINC**
Background
FAQ
Users Guide

**Use**
**LOINC**
Download LOINC
Map to LOINC
Search LOINC

**Get**
**Involved**
Forum
Meetings
Mailing Lists

**Develop**
**LOINC**
Submit Term Requests
What's Coming
Translate LOINC

**Current Versions**

**LOINC 2.38**
Released: 2011-12-30

**RELMA 5.5**
Released: 2011-12-30

Download Now

**Recent Forum Posts**

Microbiology | Re: Microbiology Text results and LOINC

Microbiology | Re: Organism Names and LOINC?

More posts...

LOINC
**LOINC**

LOINC #LOINC allows labs and MDs to talk apples to apples via @ADVANCEforAL http://t.co/EbCv9mBj
12 days ago · reply · retweet · favorite

twitter
Join the conversation

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
3. Compact representation of clinically actionable subsets for optimal performance (clinician thinkspeed = 250msec)
   **Issue:** Using keywords and short phrases e.g., CYP2C19*2*2 as shorthand for presence of CYP2C19 homozygous variant that poorly metabolizes commonly prescribed drugs such as Clopidigrel (Plavix) enables quick visual recognition and efficient lookup by decision support rules. Currently << 1% of complete genome.

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
3. Compact representation of clinically actionable subsets for optimal performance (clinician thinkspeed = 250msec)
4. Simultaneously support for human-viewable formats (with links to interpretation) and formats interpretable by decision support rules.
   **Issue:** Volume and complexity of molecular variation data exceeds cognitive capacity even of specialists. "Genomic competence" rare among primary care providers.

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
3. Compact representation of clinically actionable subsets for optimal performance (clinician thinkspeed = 250msec)
4. Simultaneously support for human-viewable formats (with links to interpretation) and formats interpretable by decision support rules.
5. Separate primary sequence data (remain true if accurate) from clinical interpretations of them (will change with rapidly changing science)
   **Issue:** Data plus interpretation in narrative lab report document is the most common current format for transmitting results back to clinicians.

# Most common current method for delivery of DNA analysis into clinical operations



DEPARTMENT OF PATHOLOGY
Nashville, Tennessee 37232

Michael Laposata, M.D., Ph.D.
Medical Director, Diagnostic Laboratories
CLIA #44D0659066

| Name: | Sex: | Laboratory Number: | VUH#: |
|---|---|---|---|
| [redacted] | F | [redacted] | [redacted] |

**Referral Source:** Dr. Kim Ely

**Reason for Request:** DNA Analysis for *KRAS* Mutations

**Type of Specimen:** Paraffin-Embedded Tissue (Block #: [redacted])

**Date Received:** 2/12/10

**Date of Report:** 2/18/10

**Interpretation:** *KRAS* Mutation NOT Detected

Mutations Tested Include:
G12A, G12C, G12D, G12R, G12S, G12V, G13C, G13D

The *KRAS* gene (12p12) is a member of the Ras family of proto-oncogenes, and encodes a protein containing guanosine nucleotide triphosphate hydrolysis activity (known more commonly as a GTPase). These proteins are active when bound to guanosine triphosphate (GTP) and inactive when bound to guanosine diphosphate (GDP). *KRAS* is membrane bound, is activated by growth factor receptors, and through BRAF, stimulates the MAPK/ERK pathway resulting in transcription and cell proliferation. *KRAS* mutations are observed in colon cancer (40-50%), lung cancer (20-30%) and pancreatic cancers (90%). Conserved missense mutations in codons 12 and 13 result in prolonged binding of GTP and constitutive activation of RAS proteins, thereby leading to uncontrolled cell proliferation.

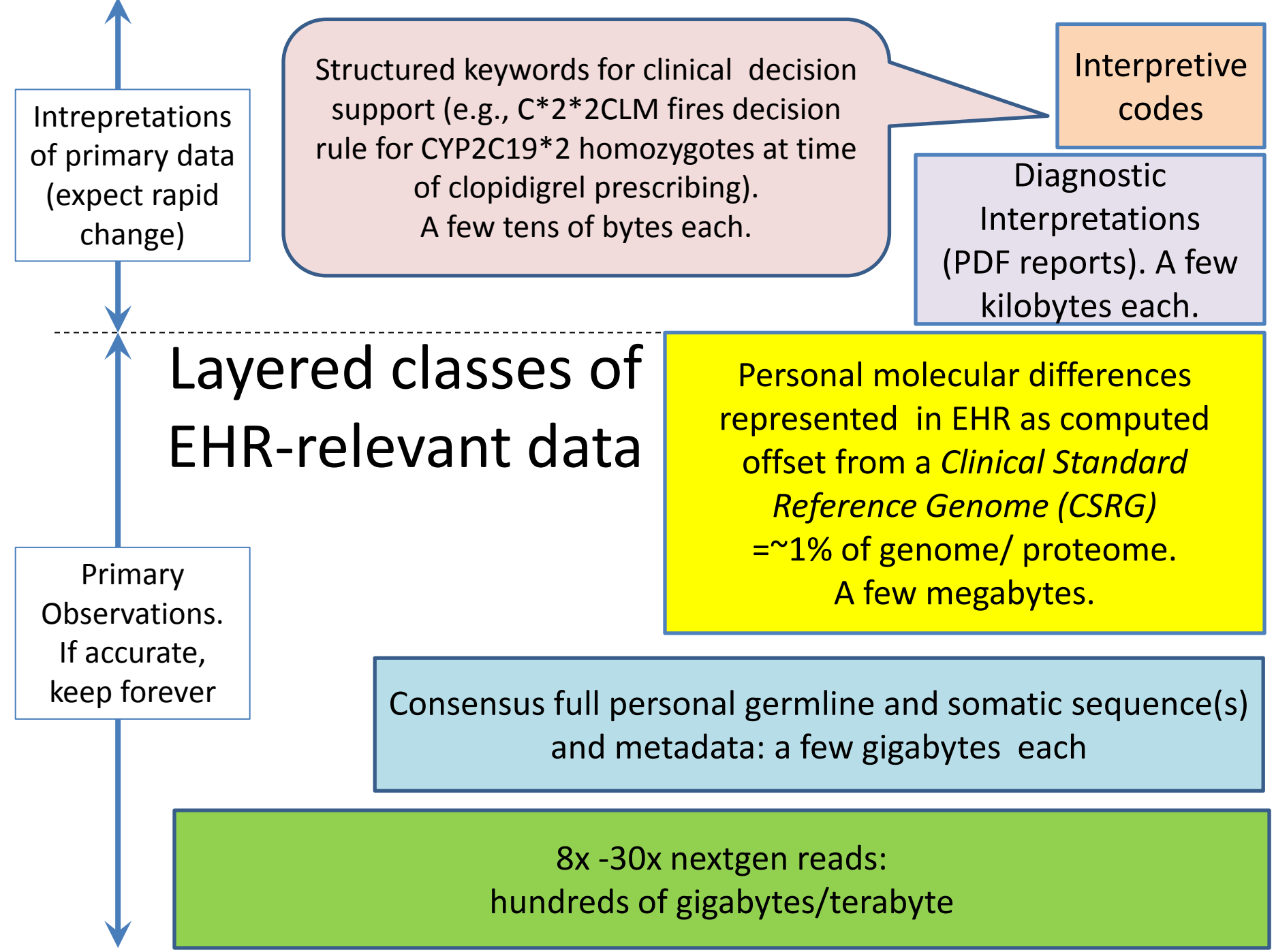
Progressive and/or metastatic non-small cell lung adenocarcinomas are often treated with inhibitors of the EGFR receptor as a second line therapy. However, it has been shown that tumors, which harbor mutations in codons

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
3. Compact representation of clinically actionable subsets for optimal performance (clinician thinkspeed = 250msec)
4. Simultaneously support for human-viewable formats (with links to interpretation) and formats interpretable by decision support rules.
5. Separate primary sequence data (remain true if accurate) from clinical interpretations of them (will change with rapidly changing science)
6. Anticipate the boundless creativity of Nature: multiple somatic genomes, multiple germline genomes for each individual over their lifetime.

# 7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them (LOINC model)
3. Compact representation of clinically actionable subsets for optimal performance (clinician thinkspeed = 250msec)
4. Simultaneously support for human-viewable formats (with links to interpretation) and formats interpretable by decision support rules.
5. Separate primary sequence data (remain true if accurate) from clinical interpretations of them (will change with rapidly changing science)
6. Anticipate the boundless creativity of Nature: multiple somatic genomes, multiple germline genomes for each individual over their lifetime.
7. Support both individual care and discovery science


Intrepretations of primary data (expect rapid change)
Structured keywords for clinical decision support (e.g., C*2*2CLM fires decision rule for CYP2C19*2 homozygotes at time of clopidigrel prescribing).
A few tens of bytes each.
Interpretive codes
Diagnostic Interpretations (PDF reports). A few kilobytes each.
Layered classes of EHR-relevant data
Personal molecular differences represented in EHR as computed offset from a *Clinical Standard Reference Genome (CSRG)*
=~1% of genome/ proteome.
A few megabytes.
Primary Observations. If accurate, keep forever
Consensus full personal germline and somatic sequence(s) and metadata: a few gigabytes each
8x -30x nextgen reads:
hundreds of gigabytes/terabyte

# Characteristics of a “Clinical Standard Reference Genome (CSRG)”

- Should generate the smallest number of differences to be stored in the EHR. (Simplest approach: most common allele for all known genes)
- Should *not* represent any actual person, ethnic group, family ancestry to avoid arguments based on those characteristics
- Should be accompanied by decompression/recompression utilities to rapidly transform any given individual’s genome or subset of it.
- Subject to version control: expect many different CSRGs as sequencing technology improves.

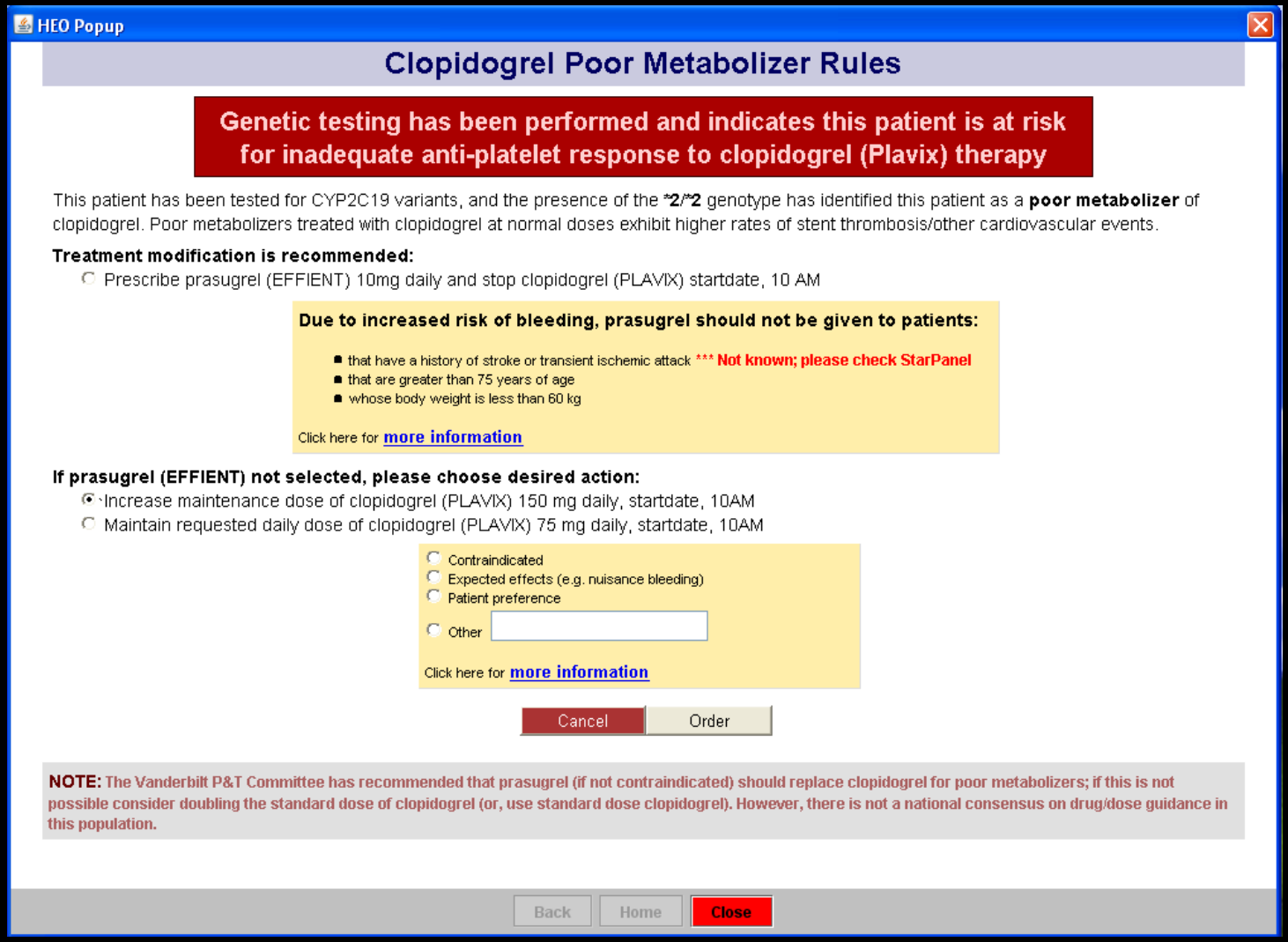

Putting it all together in an operational healthcare prototype

# PREDICT: Pharmacogenomic Resource for Enhanced Decisions in Care and Therapy

(Go-live date: Sept 15, 2010)

- Use data mining methods in Electronic Medical Record (EMR) to identify individuals at increased likelihood of a future prescription of a drug for which pharmacogenetics has relevance
- Prospectively acquire 200 marker SNP panel and put selected subset of data in electronic medical record
- At moment of prescribing, use decision support rules that look for presence of pharmacogenetic ‘keywords’ e.g., ‘CYP2C19xxxx’ to guide drug selection and correct dosing.
- Track outcomes

# Point of care decision support



# Clinician display

Alert **Clinical Trial Participant** Actions

**StarTracker Conditions/Diseases:** No Tracked Conditions ♦Customize

**** notation indicates test is due for repeat and value may be outdated.*

| Preventive | BP | BMI | eGFR | HCT | FLUVAX | CRC | Mammogram | PAP |
|---|---|---|---|---|---|---|---|---|
| | 143/72 | 31.7 ( 12/30/2010 ) | 52 | 33 | NONE | NONE | UNKNOWN | UNKNOWN |
| | SMOKE | | | | | | | |
| | UNKNOWN | | | | | | | |

**Patient-specific guidelines** MedicationsLog ICD9 History **Update** **Update (free text)** NoChang

**General Information:**

PCP:
Primary cardiologist:

**Significant Medical Diagnoses and Conditions:**

1. Coronary atherosclerotic heart disease
a. NonSTMI 01/2010
b. Coronary intervention 1/12/2010
(1) Xience 3x23 drug eluting stent to RCA
c. Coronary intervention 2/17/2010
(1) two 2.5x28 and 2.25x18 Cypher DES to LAD and diagonal
d. Coronary intervention 4/6/2010

**Adverse and Allergic Drug Reactions:**
penicillin (class) (rash)
cephalexin (rash)

**Drug Genome Interactions:** *(12/21/10 08:02)*
clopidogrel sensitivity: POOR METABOLIZER, REDUCED ANTI-PLATELET EFFECT - gene: CYP2C19 - gene result: *2/*2

**Medications:** prepare to print print and give pt. Show Hx of medications
Drug/Herb Interactions
aspirin 325 mg orally once daily, in the morning
prasugrel (effient) 10 mg orally once daily
carvedilol 6.25 mg orally twice daily with meals
lisinopril 10 mg orally once daily
furosemide 40 mg orally once daily

My Health At Vanderbilt Hom...

VanderbiltHealth.com: For Patients and Visitors

You are currently logged into MyHealthAtVanderbilt

Log Out Here

# My Health *at* Vanderbilt

Help: Using MyHealth

**Home Page For: QAPATIENTE, GREEN** beth.dunaway@vanderbilt.edu

## All Health Topics

- View all Health Topics

## Alzheimer's Disease

- What is Alzheimer's Disease?
- Alzheimer's Caregivers
- Understanding Asthma

[more stories]

## Be Healthy, Stay Healthy

- Abdominal Aortic Aneurysm Screening
- Blood Pressure Screening
- Breast Cancer Screening

[more stories]

## Health News You Can Use

- Safety Experts at Monroe Carell Jr. Childrens Hospital at Vanderbilt Warn Parents and Caregivers Against the Dangers Associated with Drop-Side Cribs
- Five Holiday Resolutions for Cancer

## Account Access

- View others who have access to QAPATIENTE, GREEN

## Messaging

- Message Your Doctor's Office

## Clinical Record

- View Your Clinical Record
- Genes that affect my medicines

## Appointments

Request an appointment with

- Your Doctor
- New Doctor
- View Existing Appointments

## Health Screenings

- Peripheral Artery Disease--are you at risk?

## Research Opportunities

- Interested in participating in research studies? Join ResearchMatch now!
- Search currently enrolling clinical studies

## Patient Resources

- Vanderbilt Giving
- Information about a

My Health At Vanderbilt Gene...

VanderbiltHealth.com: For Patients and Visitors

You are currently logged into MyHealthAtVanderbilt

Log Out Here

# My Health *at* Vanderbilt

Help: Using MyHealth

Home | Account Access | Messaging | Clinical Record | Appointments | Billing | Account Maintenance

## Genes that affect my medicines QAPATIENTE, GREEN

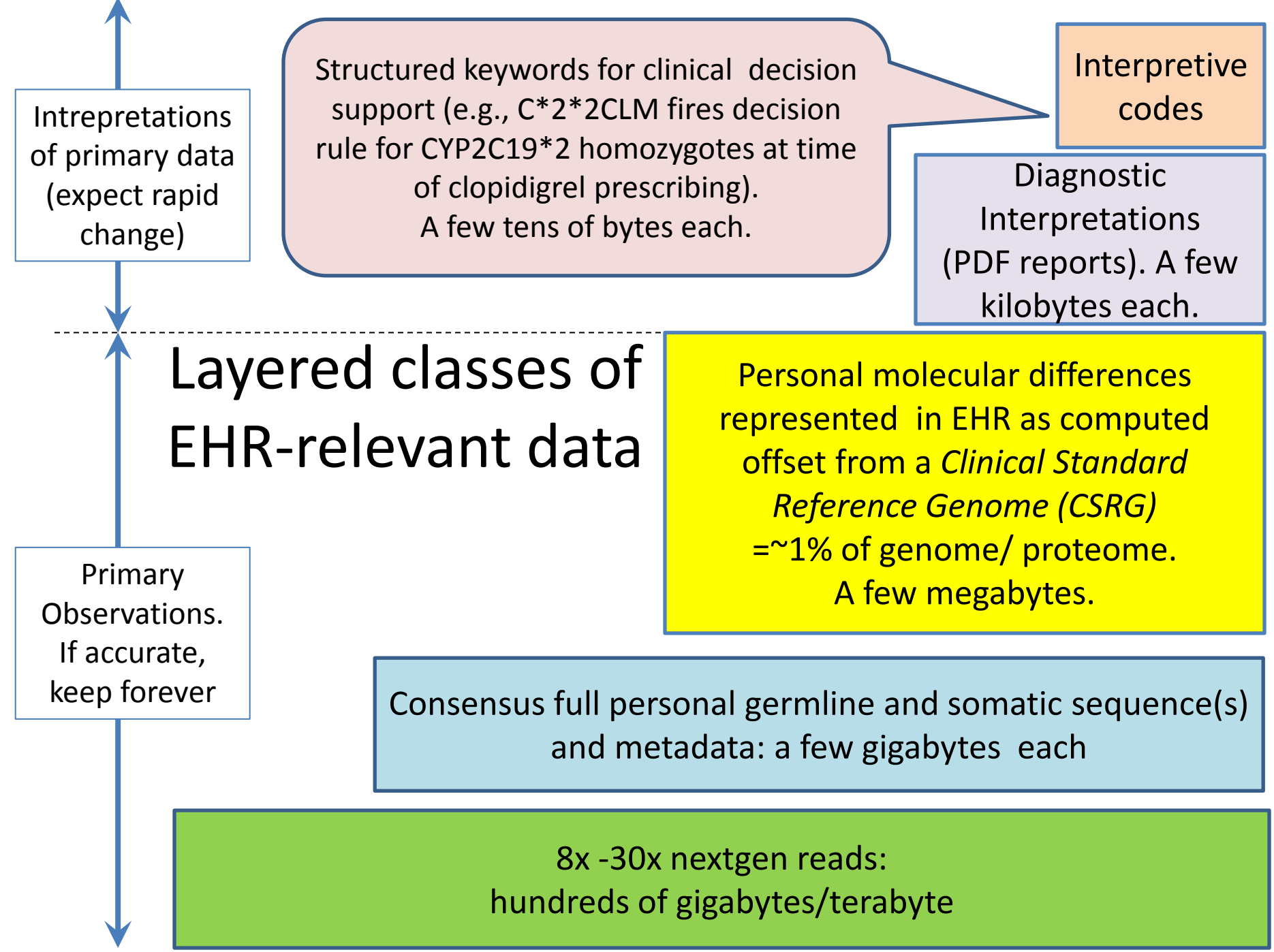
beth.dunaway@vanderbilt.edu

This test examines your gene known as CYP2C19 (sounds like "sip-2-C-19"). CYP2C19 can affect your response to a drug called clopidogrel (sounds like "kloh-PID-oh-grel"). Clopidogrel has the brand name Plavix. Clopidogrel is used to help prevent harmful blood clots from developing, such as for people who have had a recent heart attack, or a stroke.

CYP2C19 Results: Your result is *1/*2. This means you may not respond as well to clopidogrel.

Many factors, including this test, help your doctor decide if taking clopidogrel is right for you.

# The face of personalized medicine


Intrepretations of primary data (expect rapid change)
Structured keywords for clinical decision support (e.g., C*2*2CLM fires decision rule for CYP2C19*2 homozygotes at time of clopidigrel prescribing).
A few tens of bytes each.
Interpretive codes
Diagnostic Interpretations (PDF reports). A few kilobytes each.
Layered classes of EHR-relevant data
Personal molecular differences represented in EHR as computed offset from a *Clinical Standard Reference Genome (CSRG)*
=~1% of genome/ proteome.
A few megabytes.
Primary Observations. If accurate, keep forever
Consensus full personal germline and somatic sequence(s) and metadata: a few gigabytes each
8x -30x nextgen reads:
hundreds of gigabytes/terabyte